Energy vs QoE Tradeoff of Dense Mobile Networks

George Arvanitakis, Florian Kaltenberger
Eurecom Sophia Antipolis, France
{George.Arvanitakis, Florian.Kaltenberger}@eurecom.fr

***Abstract*—Energy consumption is one of the primary concerns of modern dense small cell networks. One of the key concepts to improve the energy efficiency of a dense network is to turn off a part of its base stations (BS) when they are idle or only lightly loaded, since even then a considerable amount of energy is consumed. In this paper we analytically investigate the tradeoff between energy efficiency and user experience, which is measured in terms of users' delay assuming a non-saturated traffic model. We provide an overall network performance with respect to the BS density and insights that can be valuable in terms of network design. Our analysis is based on a) a BS's linear energy consumption model b) stochastic geometry to model the topology of the network and the users and c) queuing theory in order to capture the flow-level performance. Our model is being applied to the popular LTE radio access technology but it can easily be extend to others. Our results provide guidelines and bounds that are able to predict the energy efficiency of the designed network.**

***Index Terms*—Stochastic Geometry; Queueing; LTE; Performance Analysis; Flow-level; Green Communications; Energy Efficiency, Densification.**

I. Introduction

One way to increase area capacity in cellular networks is to add more base stations (BS) a process also known as densification. Especially the addition of small cells is expected to be one of the key solutions to tackle the exponential increase of traffic on the upcoming years [1]–[3]. On the other hand, as the data traffic and the density of the networks are increasing, the energy consumption is becoming more and more crucial for both environmental (reducing of the carbon footprint) and economical reasons [4].

Studies have shown that around of 50%-70% of the total power consumption of telecommunications is taking place on the BSs [5]. A considerable amount of energy is consumed on the BS (for staying on or cooling) despite serving little or no traffic [6]. Furthermore, we should consider industrial zones, shopping roads, etc. those dense networks are deployed in order to serve the high amount of connected users on rush hours but for the rest of the day the network becomes under-utilized. Therefore, one of the key concepts for energy reduction is to turn off (or put in sleep mode) such BSs.

Depending on the radio access technology used, this can be achieved by various sleeping techniques [7]. In 3GPP LTE-Advanced (release 10) for example, the carrier aggregation feature can be used to steer traffic to another cell and to power off cells with no traffic. This feature has been improved in release 12 using the discovery reference signal, which is sent by sleeping cells only in configurable intervals [8].

The problem of switching on/off BSs has been investigated in various works. In [9] authors solve the optimization problem of user association taking into account the energy consumption as well as the flow-level performance (delay), but this analysis does not provide analytical results for the overall network performance. In [10] authors take a queuing analytic approach to study the impact of turning off a BS on neighboring ones for different traffic models, but they only do a network wise performance analysis through numerical simulations. In [11] authors follow a stochastic geometry approach in order to provide the network performance while turning off BSs, this work assumes saturated BSs and does not provide impact about the flow-level performance of the network.

Our analysis is based on a common used energy cost model [6] combined with our recently developed framework that analyzes the flow-level performance of a random placed network [12], [13]. We combine tools from stochastic geometry and queuing theory in order to analyze the whole network performance and provide insights about the tradeoff between users' QoE (which is measured in terms of delay) and energy efficiency, without assuming saturated BSs.

We apply our results for the LTE radio access technology and mainly to the case of decreasing the network density by turning off (sleep) a part of the network, but the same framework can analyze the case of increasing the network density by adding BS (densify). In oder to provide close form expressions and to avoid complex system-level simulators we assume that BSs to be turned off are selected randomly, but as we will see this worst case approach is not so far from more sophisticated criteria such as minimum associated users.

Summarizing, our contributions are:

- We derive analytical and semi analytical formulas to study the tradeoff between energy efficiency and users' delay. to the best of the authors knowledge, this is the first analytical work that combines the flow-level performance of a non-saturated network with energy consumption.There are cases where it is possible to have large energy gain and on the other hand affordable reduction of users' performance.
- We validate our analytical model by comparison with the results of packet-level simulator.
- We compare our assumption were we chose at random the BS that will be turned off with simulations of more sophisticated criteria such as minimum number of asso-

This work was funded partially by SOLDER of the EC Seventh Framework (FP7 CNECT-ICT-619687).

ciated users and we will see that both approaches are converging as the network density increases.
- In off-peak hours the amount of users is significantly reduced and the network becomes under-utilized. For this scenario, we derive the maximum amount of BS that can be switched off without affecting the performance of the remaining users. Additionally, we provide a simple rule under which conditions this BS reduction leads to energy gain (is possible that the energy consumption increases despite turning off some BSs).

The rest of the paper is organized as follows Section II presents our system model, including all of our assumptions on the topology, propagation model, scheduler etc. Section III derives the MCS distribution for the arbitrary BS. Section IV presents the BS's energy cost model and modifies it to the more useful metric, energy per unit area, the scenario of users' density reduction and some theoretical results are presented as well. Finally, in Section V we present some interesting results of our analysis about the flow-level performance and the energy efficiency of the network.

II. Our Model

A. PHY Layer Modeling

The first step is to define the assumptions about the topology (both BS and users) as well as the PHY-layer characteristics.

A.1: Both BSs and users follow a homogeneous Poisson Point Processes (PPP) with densities λ_{BS} and λ_u accordingly. Therefore, the number of BSs (or users) in an area S is

$$P(N = n \mid S) = \frac{(\lambda_{\text{BS}} S)^n e^{-\lambda_{\text{BS}} S}}{n!}, \quad n = 0, 1, \dots \ , \qquad (1)$$

and their placement is random.

A.2: A standard power loss propagation model is used, usually the path loss exponent is $2 < \alpha < 5$. Additionally, assume a Rayleigh fading channel with mean 1 and constant transmit power of P_{tx}. So, the received power at distance d from the BS is given by $P_{\text{rx}} = hd^{-\alpha}$ where h follows an exponential distribution, $h \sim \exp(P_{\text{tx}})$. Hence, the SINR if the user is associated with the i-th BS is given by

$$SINR_i = \frac{P_{\text{rx}_i}}{\sum\limits_{n \neq i} P_{\text{rx}_n} + \sigma^2} \ , \qquad (2)$$

where σ^2 is the thermal noise. Usually, $\sigma^2_{\text{dBm}} = -174 + 10\log_{10}(BW)$, where BW is the systems bandwidth [14].

A.3: We assume that all BSs have equal transmit power and implement the same scheduling policy. Additionally, we assume that each user gets connected to the closest BS, so, the BSs coverage area could be represented by Voronoi Regions (Tessellations).

Taking into account *A.1* and *A.3* the users' cardinality n for an arbitrary BS, is a *random* variable. Observe that the size of an arbitrary cell is a random variable, depending on the random BS topology, and the number of users given a specific cell size is also a random variable. The following lemma provides the probability mass function (pmf) $f_N(n)$, of users cardinality on an arbitrary cell. The proof of it as well as a useful and accurate approximation could be found at our technical report [15] or [16], [17].

Lemma 2.1: Consider BSs distributed in 2D as a homogeneous PPP with density λ_{BS}, and offering coverage to a set of users distributed as another PPP with density λ_{u}, (*A.1*). Assume further that user association within this tier is done using the closest-distance rule, (*A.3*). Then, the probability of having exactly n users in an arbitrary cell, $f_N(n)$, is given by:

$$f_N(n) = \frac{343}{n!15}\sqrt{\frac{7}{2\pi}}\frac{\zeta^n}{(\zeta + \frac{7}{2})^{n+\frac{7}{2}}}\Gamma(n + \frac{7}{2}) \ , \qquad (3)$$

where $\zeta = \frac{\lambda_{\text{u}}}{\lambda_{\text{BS}}}$ and Γ is the gamma distribution.

A.4: In this work we assume that the part of the network that will be turned off is selected randomly. The random selection can serve as a worst case scenario (assuming that we do not exploit our knowledge of the network in order to intentionally decrease its performance). There are more sophisticated criteria that can decide which BSs to turn off in order to improve the energy performance of the network (minimum amount of connected users, minimum providing service rate, etc.) but those criteria, in most cases cannot be modeled mathematically.

If we turn off randomly the 10% of the initial BSs, the remaining are again a Poisson process with densisty BS, $\lambda^{'}_{BS} = 0.9\lambda_{BS}$ due to the following lemma.

Lemma 2.2: Let a Poisson process with rate λ if we divide it randomly with probability p and $(1-p)$ to two processes, Then, the two outcome processes are again Poisson with new rates $\lambda^{'}_1 = p\lambda$ and $\lambda^{'}_2 = (p-1)\lambda$ respectively due to Poisson thinning property.

B. BS level Modeling

We assume that each BS experiences a *dynamic* traffic load and we would like to study the performance at *flow-level*. We now state our assumptions regarding a single randomly chosen BS, and comment where necessary.

A.5: Each *connected* user to a BS generates new *flow* requests randomly, and independently of other users, according to a Poisson Process with density λ_f.

A.6: A flow is a sequence of packets corresponding to the same user or application request (e.g., a file or web page download). Each flow has a random size, in terms of bits, drawn from a *generic* distribution with mean value $\langle s \rangle$.

The following Lemma follows easily, by using a simple Poisson merging argument [18].

Lemma 2.3: If n users are associated with a given BS, the aggregate flow arrival process to that BS is Poisson$(n\lambda_f)$.

Remark: While a Poisson arrival model is pretty standard in related literature, note that if the number of users n at a BS is relatively large, assumption (A.1) can be relaxed to more general traffic arrivals, and we can then use the Palm-Khintchine theorem [18] to support Lemma 2.3 as an approximation.

A.7: In the absence of other flows, *a single flow will be served at full rate*, with the maximum Modulation and Coding

Scheme (MCS) that the BS can offer to that user, which in turns depends on the SINR-BLER specifications for that RAT. In this work we examine an LTE network, so, according to [16] the SINR thresholds and the corresponding rate for each MCS depicted in table I.

TABLE I: LTE's SINR threshold (τ) in dB and the corresponding rate (MB/s) w.r.t. MCS index, for the case of 20MHz bandwidth and acceptable BLER 10^{-1}

#	τ	rate	#	τ	rate	#	τ	rate
0	-2.3	2.8	9	3.8	15.8	18	10.3	32.9
1	-1.6	3.6	10	5.3	16.0	19	11.5	36.7
2	-1.0	4.6	11	5.5	17.6	20	12.1	39.2
3	-0.2	5.7	12	5.9	19.8	21	12.9	43.8
4	0.6	7.2	13	6.8	22.9	22	13.4	46.9
5	1.3	8.8	14	7.9	25.5	23	14.6	51.0
6	1.8	10.3	15	8.6	28.3	24	15.3	55.1
7	2.6	12.2	16	9.1	30.6	25	16.0	57.3
8	3.2	14.1	17	10.2	30.8	26	16.9	61.7

The MCS probability mass function $f_{\text{MCS}}(mcs)$ is derived in Section III.

We will assume a SISO system and a single carrier in our analysis [19]. Increased rates due to spatial multiplexing and carrier aggregation could be included with a proper physical abstraction models.

C. Queueing Model for BS Schedulers

When more than one flows are served in parallel by a BS, the BS operates as a *queueing system*. The service rate for a flow is generally smaller than what assumption (A.6) predicts. It actually depends on the number of active flows (BS load), and the BS scheduler or media access control (MAC) mechanisms which decide how the available resources will be distributed between flows.

Resource Fair Scheduler: Assume all flows are allocated the same amount of resources by the BS, and are served simultaneously, e.g., in a round-robin, TDMA-like manner. If the service time slot is small (e.g., of packet size) compared to the total size of a flow, the flow level performance at that BS can be approximated by a multi-class M/G/1 Processor Sharing (PS) system. This model has already been used to analyzed 3G/3G+ BS performance [20], [21].

LTE schedulers are significantly more complex, allocating competing flows both time and frequency resources (Resource Blocks), possibly taking into account the queue backlog of each flow and flow priority, and also attempting to take advantage of instantaneous SINR variations in time and frequency to achieve further multi-user diversity [19]. While a large number of algorithms have been proposed [22], in the lack flow priority, most implemented schedulers lead to a proportionally fair throughput allocation between flows [19].

The following is a direct application of the multi-class M/G/1/PS result [23].

Lemma 2.4: For a BS with n users generating flows of mean size $\langle s \rangle$, with instantaneous transmission rates $r(mcs)$ drawn from distribution $f_{\text{MCS}}(mcs)$, and allocated resources by a resource fair scheduler, the effective service rate of the cell is

$$\langle \mu \rangle = \left(\sum_{mcs} \frac{f_{MCS}(mcs) \cdot \langle s \rangle}{r(mcs)} \right)^{-1} \text{ flows/sec}, \quad (4)$$

and the mean flow delay is given by

$$Delay = \frac{1}{\langle \mu \rangle - n\lambda_f} \ , \quad (5)$$

The load of a system could be defined as $\rho = \frac{\text{input job rate}}{\text{service job rate}}$, for our case the average network load could be defined as

$$\rho = \frac{\zeta \cdot \lambda_f}{\langle \mu \rangle} \ , \quad (6)$$

where $\zeta = \lambda_u / \lambda_{BS}$, additionally, when the system is stable $\rho < 1$. There are a lot of ways to express network load, we chose the following, the percentage of time that a BS is ON. Performance gains from opportunistic scheduling can be included in the above equation as a multiplicative factor in front of $\langle \mu \rangle$. Additionally, another kind of schedulers could be modeled with different queues i.e. WiFi scheduler [24].

III. MCS Distribution

A. Coverage Probability

The probability that a user's SINR is grater than a given threshold τ, $P(SINR > \tau)$, is called coverage probability. The coverage probability for the arbitrary user in a random placed cellular network, assuming that the surrounding BSs are interfering only for the amount of time that are serving a user has been derived in [12],

$$p_c^{lb}(\tau, \alpha, N_{max}) = \sum_{n=0}^{N_{max}-1} \left(f_N(n \mid \zeta) \frac{1}{1+\mathcal{A}_\rho} \right) + \overline{F_N}(N_{max} \mid \zeta) \frac{1}{1+\mathcal{A}_{\rho=1}} \ . \quad (7)$$

Where $N_{max} = \langle \mu \rangle / \lambda_f$ is the maximum amount of associated users per BS, $\mathcal{A}_\rho = (\tau\rho)^{2/\alpha} \int_{(T\rho)^{2/\alpha}}^{\infty} \frac{1}{1+u^{\alpha/2}} du$, we can consider the BS load as $\rho = \frac{n}{N_{max}}$. Additionally, α is the path loss exponent, f_N and $\overline{F_N}$ are the pdf and ccdf of users' cardinality and as previously used $\zeta = \lambda_u / \lambda_{BS}$, (see Lemma 2.1). Further simplifications of Eq. (7) could be found in [12].

B. MCS distribution

Assuming an SINR threshold τ_i for each MCS (table I), the pmf of the MCS $f_{MCS}(mcs_i)$ can be obtained through the coverage probability as follows

$$f_{\text{MCS}}(mcs_i) = p_c(\tau_i, \alpha, N_{max}) - p_c\left(\tau_{(i+1)}, \alpha, N_{max}\right) \ . \quad (8)$$

Taking into account that $N_{max} = \langle \mu \rangle / \lambda_f$, we can observe from Eq. (7) that the coverage probability depends on service rate $\langle \mu \rangle$. Thus, MCS distribution depends on the $\langle \mu \rangle$ as well. On the other hand, $\langle \mu \rangle$ is depending on MCS distribution as we can see at Eq. (4).

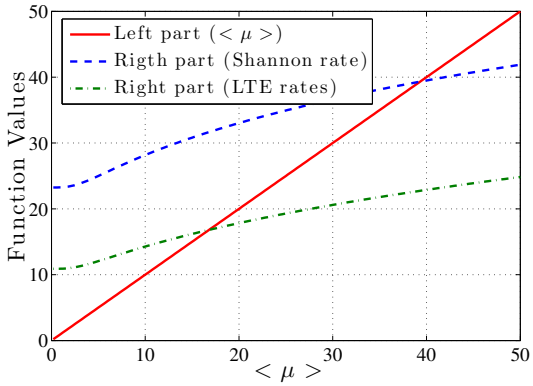

Fig. 1: Left and right part of the Eq. (9)

C. Service Rate $\langle\mu\rangle$

Due to the aforementioned dependencies we re-write Eq. (4) to more proper form

$$\langle\mu\rangle = \left(\sum_{mcs_i} \frac{f_R(mcs_i|\,\langle\mu\rangle)\cdot\langle s\rangle}{r(mcs_i)}\right)^{-1} . \tag{9}$$

With respect to $\langle\mu\rangle$ Eq. (9) has exactly one solution. As a remark we mention that the left part of equation Eq. (9) is a strictly increasing function with derivative equal to one, and the right part is again a strictly increasing function with respect to $\langle\mu\rangle$ but its derivative is strictly smaller than one for the cases of LTE or Shannon's rates (calculated computationally). Fig 1 depicts the left and the right part of Eq. (9) w.r.t. $\langle\mu\rangle$ for two cases a) the right part of the equation computed assuming LTE rate and b) assuming Shannon's rate. In both cases there is exactly one solution for the Eq. (9) and could be approached by simple gradient methods.

IV. Energy Cost Model

The linear cost model is the most common energy consumption model [6]. The model consists of: a) a constant term that captures energy consumption of the BS in order to be ON and ready to operate and b) a linear term that is responsible for the energy consumption while the BS is operating (exchanging data). We set as E_{on} the amount of energy that the BS consumes in order to be ON for a period T. Additionally, E_{op} is the energy consumption while the BS is operating with a user and Δt is the amount of time that the BS is operating ($\Delta t \leqslant T$). So, the total energy consumption of the BS for period T is given by

$$E_{BS} = E_{on} + E_{op}\frac{\Delta t}{T} . \tag{10}$$

We are not interested about the energy consumption of a specific BS but for the total energy consumption per unit area. After some trivial calculations and expressing $\frac{\Delta t}{T}$ as network load ρ we end up with

$$\overline{E} = \lambda_{BS}\left[E_{on} + E_{op}\cdot\rho_{(\zeta)}\right] , \tag{11}$$

where $\rho(\zeta)$ is the average utility of the network Eq. (6) and $\zeta = \frac{\lambda_u}{\lambda_{BS}}$.

Partial Reduction of Users Density (Off-Peak Hours)

A network is deployed in order to achieve a specific user performance. Some areas suffers for high users' variability in a day, shopping streets, industrial zones, etc. The network design aims to achieve a predefined users' performance even at the high traffic period. That means that the network is under-utilized on the low traffic period. The operator's goal is to turn off a part of the network in order to save energy, but without decreasing the performance of the remaining users.

Lemma 4.1: If we assume that the density of the users in a given area decreased according to a factor l_u and the thermal noise is negligible compared to interference $\sigma^2 << I$, the maximal decrease factor of the BS density l_{bs} but without affecting the average delay of the remaining users is

$$l_{bs} = l_u . \tag{12}$$

Proof: The Delay in both cases (before and after users reduction) should be the same

$$Delay_1 = Delay_2 ,$$

where according to processor sharing

$$\frac{1}{\langle\mu\rangle_1 - \lambda_1} = \frac{1}{\langle\mu\rangle_2 - \lambda_2} ,$$

were we express service rate $\langle\mu\rangle$ as a function of the densities ratio $\zeta = \frac{\lambda_u}{\lambda_{BS}}$. Assuming that the other network parameters constant

$$\left\langle\mu_{(\zeta)}\right\rangle - \zeta\cdot\lambda_f = \left\langle\mu_{(\frac{l_u}{l_{bs}}\zeta)}\right\rangle - \frac{l_u}{l_{bs}}\zeta\cdot\lambda_f . \tag{13}$$

further we define

$$g(x) = \left\langle\mu_{(x)}\right\rangle - x\cdot\lambda_f . \tag{14}$$

Both $\left\langle\mu_{(x)}\right\rangle$ and $-x\cdot\lambda_f$ are strictly decreasing functions with respect to x. So $g(x)$ is a strictly decreasing function as well. If $g(\cdot)$ is a strictly monotonic function and $g(a) = g(b)$ then $a = b$. Thus, applying that to Eq. (13) we have

$$\zeta = \frac{l_u}{l_{bs}}\zeta . \tag{15}$$

■

Unfortunately, switching off some BS does not lead directly to energy improvement. When we switch off a part of the network, on the one hand we save some energy by E_{on} factor of the BSs that we turned off, but on the other hand, the network in total consumes higher amount of E_{op} than before, because the remaining BSs serve the users with worse channel conditions (therefore, for more time) than before.

Lemma 4.2: The BS reduction according to the reduction of users density that was presented in Lemma 4.1, surely reduces the energy consumption of the network if

$$\frac{E_{on}}{E_{op}} > \frac{1}{1-l_{bs}} . \tag{16}$$

Proof: Regarding the energy consumption of this scenario the answer is not trivial. We aim the energy consumption of the thinned $\overline{E}_{th}$ network to be less than the consumption of the full network $\overline{E}_f$

$$\overline{E}_{th} < \overline{E}_f \ . \tag{17}$$

Assuming that for a portion of time p_t the users density has been reduced by a factor l_u, thus, according to Lemma 4.1 the BS's density will reduced by a factor $l_{BS} = l_u$ and by applying to the Eq. (11)

$$p_t l_{BS} \lambda_{BS}[E_{on} + \rho_{th} E_{op}] < p_t \lambda_{BS}[E_{on} + \rho_f E_{op}] \ , \tag{18}$$

after some trivial calculations we derive that in order to have some energy gain through the thinning of BS the following inequality should hold

$$\frac{E_{on}}{E_{op}} > \frac{l_{bs}\rho_{th} - \rho_f}{1 - l_{bs}} \ . \tag{19}$$

taking into account that the nominator of Eq. (19) is upper bounded by $l_{BS}\rho_{th} - \rho_f < 1$, we end up to the asymptotic rule of thumb of Eq. (16) ■

If Eq. (16) does not hold, we should calculate the load for each case by following the whole procedure of Section III and Eq. (6) as to decide if the energy consumption of the network will be decreased by turning off a part of the network or not.

V. Results / Sensitivity analysis

A. Validation

The packet-level simulator generates both BSs and users randomly placed in a large surface with given densities (λ_{BS}, λ_u). Users are associated with the closest BS and generate flows according to Poisson distribution with density λ_f. The flows are forwarded to the corresponding BS which is modeled as a multi-class M/G/1/PS. The service rate of each flow for every time quantum is calculated via SINR. At the calculation of the interference we are taking into consideration only the base stations that are ON at this time quantum (load based case), for comparison with the most common assumption in the literature we also implement the case of taking into consideration, at the interference calculation, all the BS (saturated case). In order to compare fairly both interference cases we consider only the users whose SINR is at least higher than the threshold of the lowest MCS at saturated case. We should mention that the packet-level simulator needs extremely high computational resources, so, the theoretical prediction is even more valuable.

We are interested in investigating the scalable performance of a large, random placed network while we turn off a percentage of BSs in order to improve the energy efficiency. In this section the *user's density does not change*, so by turning off BSs we know a priori that users' performance will be reduced, so we want to compare the energy gain with the performance reduction. In our theoretical analysis we select randomly which BSs will turn off. As we mention before, the random selection is a worst case scenario, there are more

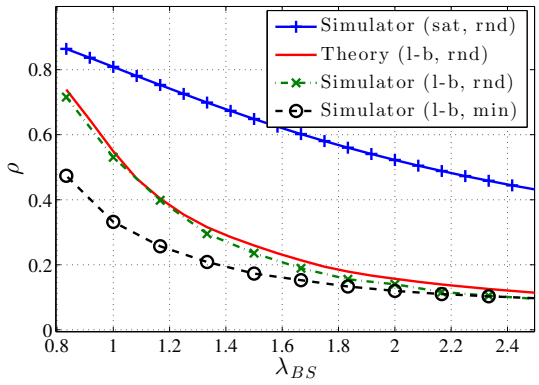

Fig. 2: Average network load, theoretical vs simulation results

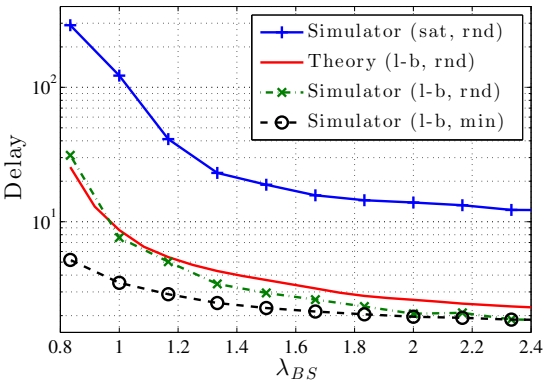

Fig. 3: Average flow delay, theoretical vs simulation results

sophisticated criteria in order to decide which BSs to turn off (minimum associated users).

Figs. 2 and 3 shows how the flow level performance of the network (average flow delay and the average network load) scales with respect to the density of the BS (λ_{BS}) in four different cases i) our theoretical prediction that assumes load based interference (l-b) and the selection of which BSs to turn off is random (rnd) ii) simulation results for the case of load based interference and the selection of which BSs to turn off is random iii) simulation results assuming load based interference and the selection of which BSs to turn according to minimum associated users (min) criterion and iv) simulation results assuming saturated (sat) BSs and the selection of which BSs to turn off is made randomly. We can obtain three interesting conclusions from Figs 2 and 3: 1) our theoretical prediction is very accurate (for both load and delay) compared with the simulation results, 2) the assumption of saturated BSs (which is common at stochastic geometry works) changes totally the network's performance not only quantitative but qualitative as well and 3) the minimum associated users criterion does not differ a lot from the random one, especially when the network is very dense the difference is negligible. That means that for dense networks it is better to randomly switch off BS than

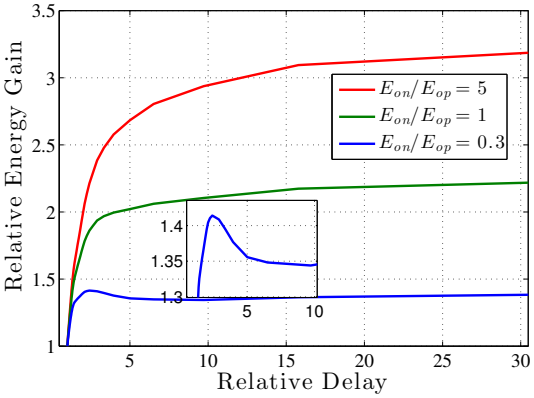

Fig. 4: Performance plots for different $\frac{E_{on}}{E_{op}}$ ratios

using the centralized and more complex minimum associated users criterion to determine which BS to switch off.

B. Energy Vs Delay

In this subsection we are interested to investigate the trade-off between energy efficiency and users delay. Regarding the interference we assume the more realistic case of load based. Initially, the network is under-utilized ($\rho \approx 0.1$), let E_0 and D_0 be the energy consumption and the average delay of this initial state. Then gradually we turn off a part of the network, we define as $\overline{E}/E_0$ the relative energy gain and as $\overline{D}/D_0$ the relative delay, for simplicity we will call those metrics energy gain and relative delay respectively.

Fig. 4 shows the performance for different $\frac{E_{on}}{E_{op}}$ ratios. Initially, by observing the case of $E_{on} = E_{op}$ we note that for low load the derivative of the energy gain is very high, so there is the capability of energy improvement without large delay cost. When the load of the network is $\rho \approx 0.5$ the derivative decreased dramatically, thus the delay cost is extremely high compered to the energy gain.

Furthermore, in Fig. 4 there two more remarks 1) as the ratio between constant and operational energy becomes higher, the possible gain by turning off the BS is increasing. Considering the traditional small cells, the case that the constant energy term be much higher than the operational one seems unrealistic, but in future networks (e.g. drones) this could be the case, 2) when the constant energy term is less than the operating one, there is a turning point in the performance curve. This means that after a point, as we turn off more BS, both the energy as well as the delay performance are getting worse. For the extreme case where the particular cost of a BS to be ON is negligible compared to the operational cost $\frac{E_{on}}{E_{op}} \to 0$ there is no capability for energy improvement, thus, the optimal strategy is simply to set all BSs ON.

VI. Conclusions / Future Work

We presented an analytical frame work that provides the energy performance of the network and we investigated the tradeoffs between network's energy efficiency and user's QoE. We observe that even under the worst case assumption of random selection of which BS turn off, there is capability of energy improvement without affecting a lot user's QoE when the operational energy in not much greater than the constant energy term. As future work, we will expand the framework to heterogenous networks and investigate different tier association criteria that maximize networks energy efficiency of the network constrained to user's QoE.

References

[1] "Strategies for mobile network capacity expansion," *Real Wireless, White Paper*, 2010.

[2] "Looking ahead to 5G," *Nokia Solutions and Networks, White Paper*, Dec. 2013.

[3] X. Ge, S. Tu, G. Mao, C. X. Wang, and T. Han, "5G ultra-dense cellular networks," *IEEE Wireless Communications*, 2016.

[4] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczok, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Communications Magazine*, 2011.

[5] M. Deruyck, W. Vereecken, E. Tanghe, W. Joseph, M. Pickavet, L. Martens, and P. Demeester, "Power consumption in wireless access network," in *Wireless Conference (EW)*, 2010.

[6] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Communications*, 2011.

[7] J. Wu, Y. Zhang, M. Zukerman, and E. K. N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE Communications Surveys Tutorials*, 2015.

[8] E. Dahlman, S. Parkvall, and J. Skold, *4G, LTE-Advanced Pro and The Road to 5G*. Academic Press, 2016.

[9] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE Journal on Selected Areas in Communications*, 2011.

[10] N. Sapountzis, S. Sarantidis, T. Spyropoulos, N. Nikaein, and U. Salim, "Reducing the energy consumption of small cell networks subject to qoe constraints," in *GLOBECOM IEEECOM*, 2014.

[11] Y. S. Soh, T. Q. S. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, 2013.

[12] G. Arvanitakis, T. Spyropoulos, and F. Kaltenberger, "An analytical model for flow-level performance of large, randomly placed small cell networks," in *GLOBECOM IEEE*, 2016.

[13] ——, "An analytical model for flow-level performance in heterogeneous wireless networks," *submited to IEEE Transactions on Wireless Communications*, 2017.

[14] A. F. Molisch, *Wireless Communications*. Wiley, 2010.

[15] G. Arvanitakis, "Distribution of the number of poisson points in poisson voronoi tessellation," Eurecom, Tech. Rep. RR-15-304, 2014.

[16] G. Arvanitakis and F. Kaltenberger, "Stochastic analysis of two-tier hetnets employing LTE and WiFi," in *EUCNC*, 2016.

[17] S. M. Yu and S. L. Kim, "Downlink capacity and base station density in cellular netwrorks," in *WiOpt*, 2013.

[18] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, 2013.

[19] S. Sesia, I. Toufik, and M. Baker, *LTE - the UMTS long term evolution: from theory to practice*. Wiley, 2009.

[20] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *Networking, IEEE/ACM Transactions on Networking*, 2005.

[21] T. Bonald and A. Proutiere, "Wireless downlink data channels: User performance and cell dimensioning," in *ACM MOBICOM*, 2003.

[22] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in lte cellular networks: Key design issues and a survey," *Communications Surveys Tutorials, IEEE*, 2013.

[23] G. Fayolle, I. Mitrani, and R. Iasnogorodski, "Sharing a processor among many job classes," *J. ACM*, 1980.

[24] G. Arvanitakis and F. Kaltenberger, "Phy layer modeling of LTE and WiFi RATs," Eurecom, Tech. Rep. RR-16-317, 2016.